Recherches sur Diderot et sur l'Encyclopédie

31-32 | Avril 2002
L'Encyclopédie en ses nouveaux atours électroniques: vices et vertus du virtuel

Lectures assistées de l'*Encyclopédie* électronique : PHILOLOGIC et WEBLEX

Serge HEIDEN and Pierre LAFON : An assisted reading of the electronic Encyclopédie : PHILOLOGIG and WEBLEX.

Serge Heiden and Pierre Lafon

Electronic version
URL: http://journals.openedition.org/rde/2913
DOI: 10.4000/rde.2913
ISSN: 1955-2416

Publisher
Société Diderot

Printed version
Date of publication: 15 April 2002
Number of pages: 91
ISSN: 0769-0886

Electronic reference
Serge Heiden and Pierre Lafon, « Lectures assistées de l'*Encyclopédie* électronique : PHILOLOGIC et WEBLEX », *Recherches sur Diderot et sur l'Encyclopédie* [Online], 31-32 | Avril 2002, Online since 03 October 2015, connection on 01 May 2019. URL : http://journals.openedition.org/rde/2913 ; DOI : 10.4000/rde.2913

Serge HEIDEN
Pierre LAFON

Lectures assistées de l'*Encyclopédie* électronique : PHILOLOGIC et WEBLEX

C'est à l'équipe d'ARTFL que revient le mérite d'avoir mis à notre disposition le texte de l'*Encyclopédie* sur le Web. S'agissant d'une œuvre aussi importante, dont l'édition papier a déjà suscité une masse de travaux érudits, l'enregistrement informatisé du texte constitue en soi une espèce de gageure. Nous saluons ici l'heureuse initiative de l'équipe de Chicago qui a eu le courage, la volonté et la ténacité nécessaires pour mener à bien cette énorme entreprise.

Mais il est vrai qu'une fois que le texte est enregistré, le problème reste entier de savoir comment appréhender ces données, avec quels instruments de lecture, de manière à bénéficier de l'aide d'une machine pour étendre et approfondir les investigations sur le texte.

Nous avons travaillé avec nos collègues Arnold et Geffroy qui, cherchant des réponses aux questions sémantiques qu'elles posent à l'*Encyclopédie*, ont eu recours à deux logiciels : PHILOLOGIC, proposé par ARTFL, qui accompagne le texte électronique et permet de le questionner et WEBLEX, conçu dans notre équipe, pour mettre en œuvre des analyses lexicométriques de corpus textuels.

Il ne s'agit en aucun cas pour nous de faire ici une évaluation comparée de ces deux logiciels qui, à beaucoup d'égards, apparaissent comme complémentaires, mais plus modestement, de proposer quelques éléments de réflexion méthodologiques sur les lectures et les recherches textuelles assistées par l'informatique. Celles-ci font l'objet des deux premières parties.

Nous consacrons la troisième partie à une petite enquête sur la fidélité typographique du texte électronique à l'édition papier, car celle-ci conditionne, bien sûr, la qualité des lectures et des recherches électroniques à venir.

Recherches sur Diderot et sur l'*Encyclopédie*, 31-32, avril 2002

1. Quelques remarques sur PHILOLOGIC

PHILOLOGIC, le programme qui permet d'interroger le texte de l'*Encyclopédie* est un programme que nous qualifierons de « documentaire ». Il permet de formuler toutes sortes de requêtes afin de retrouver des informations dans et sur le texte.

En l'absence de spécifications initiales, l'utilisateur est en relation avec la totalité du texte. Il peut restreindre les recherches à une partie du texte en spécifiant un domaine, un auteur, un/des article(s), ou une combinaison de ces trois entités.

Ce programme a de très grandes qualités : les principales, nous semble-t-il, résident dans sa simplicité d'utilisation et dans la rapidité étonnante des réponses qu'il fournit.

PHILOLOGIC constitue un instrument d'exploration remarquable pour chercher des attestations d'expressions linguistiques, pour les localiser, pour calculer leur fréquence, pour saisir leur sens dans leur environnement linguistique. Outil d'exploration du texte et d'assistance à sa lecture, PHILOLOGIC permet de fouiller, de scruter et de pénétrer peu à peu une œuvre aussi vaste, aussi complexe et aussi diverse que l'*Encyclopédie*, sans s'y perdre.

Les fonctionnalités principales de PHILOLOGIC sont bien adaptées aux interrogations en temps réel sur le Web. Sans les changer, il nous semble possible d'y apporter deux améliorations.

La première est minime. PHILOLOGIC entrouvre la porte du quantitatif puisqu'il fournit à la demande des fréquences d'occurrences. Cependant, il est difficile de faire des comparaisons entre ces fréquences et, en général, de les interpréter, car la longueur du texte sur lequel elles sont calculées n'est pas connue. Une indication chiffrée de cette longueur compléterait utilement les fréquences calculées. Il nous semble que l'utilisateur devrait pouvoir accéder à la longueur des articles et du sous-ensemble textuel sur lequel il travaille. Il pourrait ainsi faire des comparaisons relatives.

La seconde est plus importante. En travaillant à la délimitation du corpus « Hugees » (voir la contribution d'Arnold et Geffroy), la question se posait ainsi : comment trouver dans l'*Encyclopédie* tous les articles qui contiennent à la fois « humanité », « genre humain » et « espèce humaine ». Or PHILOLOGIC ne permet pas, pour le moment, de répondre directement à cette question. Il faut chercher la liste des articles qui contiennent « humanité », celle des articles qui contiennent « genre humain », celle enfin des articles qui contiennent « espèce humaine », puis faire l'intersection des trois listes.

Pour pouvoir faire une telle recherche automatiquement, il faudrait :

Soit introduire l'entité *article* dans la recherche en texte intégral. Pour l'instant PHILOLOGIC propose seulement la *phrase* et le *paragraphe*.

Soit donner à l'utilisateur la possibilité d'une construction par étapes de ses requêtes : dans un premier temps on chercherait tous les articles contenant « humanité », puis sur les articles obtenus tous ceux qui contiennent « genre humain », puis enfin sur le nouvel ensemble tous ceux qui contiennent « espèce humaine ».

Ces deux propositions convergent pour faire jouer un rôle accru à l'entité *article* dans l'architecture des requêtes de PHILOLOGIC. L'article, en effet, constitue l'unité textuelle élémentaire de l'*Encyclopédie*, et les investigations sur son texte ou sa langue, qu'elles soient sémasiologiques ou onomasiologiques, conduisent à délimiter des corpus de travail construits à partir de la sélection d'un sous-ensemble d'articles. C'est, en tous cas, la situation dans laquelle nous nous sommes trouvés, situation que nous pensons très fréquente parmi les requêtes des utilisateurs de l'édition électronique.

Pour décrire et analyser des corpus textuels de ce type, c'est-à-dire composés d'une addition de fragments, il est préférable de recourir à des logiciels conçus à cette fin. Ils permettent, notamment, d'établir des contrastes entre les diverses composantes du corpus. Nous décrivons ci-dessous, brièvement, les possibilités de l'un d'eux, WEBLEX, conçu et mis au point dans notre équipe par Serge Heiden.

2. Présentation de WEBLEX

Le logiciel WEBLEX est composé d'un ensemble de programmes mettant en œuvre la méthode d'analyse lexicométrique des textes[1]. De par son interface, liée à l'Internet, et certaines de ses fonctionnalités, il est comparable au système PHILOLOGIC tel qu'on le rencontre dans la version de l'*Encyclopédie* de l'ARTFL. Mais il est doté de fonctions complémentaires spécifiques qui enrichissent les analyses et les constats que l'on peut faire sur un corpus textuel numérisé. L'expérience d'analyse d'un sous-corpus d'articles et de sous-articles de l'*Encyclopédie* sous le nom de « Hugees » est pour nous l'occasion de montrer ces spécificités.

Il est utile de distinguer dans l'application de l'outil informatique à la représentation et à l'analyse d'un document deux aspects complémentaires et indissociables : d'une part la donnée (ou *l'objet*) *représentant* le document en tant que tel dans la machine sous forme numérique et d'autre part, les programmes (ou les *actions*) permettant d'*appréhender* cette

1. La méthode lexicométrique est née et a été élaborée progressivement dans le laboratoire de lexicologie de l'ENS de Fontenay-Saint-Cloud (actuellement UMR8503, Analyses de corpus linguistiques) depuis 1967. WEBLEX s'inscrit dans la continuité de son développement, au même titre que le logiciel LEXICO d'André Salem (Lebart L., Salem A. 1994).

donnée selon diverses modalités. Bien sûr, données et actions ont un degré de complexité très variable. Le spectre des données peut s'étendre d'une simple suite de nombres à un ensemble complexe de documents structurés, hiérarchiquement analysables à travers leurs occurrences. On comprend le lien étroit entre ces deux notions dans la mesure où l'utilisateur de l'outil ne pourra finalement exploiter que les informations présentes dans la machine. Cela concerne non seulement les modes d'accès aux données mais aussi l'exactitude de leur recueil (ou encodage). Chaque logiciel propose sa propre *interface*, entre l'utilisateur et la machine, pour réaliser cette exploitation et visualiser les résultats.

Dans le cas de l'*Encyclopédie* de l'ARTFL, l'objet est structuré en articles, sous-articles, paragraphes et occurrences graphiques, ainsi qu'en images numérisées des pages. Les actions proposées s'appliquent aux vedettes ou aux catégories des articles, ainsi qu'aux occurrences de leur contenu (la recherche « plein texte »). Elles sont du type « rechercher » puis « feuilleter ». L'interface hypertextuelle de l'Internet permet de les déclencher par une simple saisie au clavier suivie de quelques parcours de liens.

WEBLEX et la lexicométrie proposent d'appréhender l'objet empirique textuel à travers deux prismes complémentaires : une approche quantitative et une approche qualitative.

L'approche quantitative focalise l'attention sur les fréquences d'apparition des événements textuels dans un document et accorde à celles-ci une valeur de symptôme.

Les événements textuels sont de divers types : le plus simple à repérer est la graphie (token en anglais), par exemple, la chaîne de caractères « femme ». On peut aussi regrouper sous un prototype les formes « femme » (singulier) et « femmes » (pluriel), ou considérer ensemble les quatre graphies « femme », « Femme », « femmes » et « Femmes ». On peut également traiter des séquences polylexicales paraissant figées, telles que « genre humain » ou « espèce humaine » comme une seule unité. On peut enfin, à condition que le texte soit enrichi d'un étiquetage morpho-syntaxique, compter des événements correspondant à des lemmes (groupant toutes les flexions d'un même mot) ou à des schémas syntaxiques, par exemple « Nom + Adjectif » qui regrouperait toutes les séquences correspondantes telles que « genre humain », « espèce humaine » et beaucoup d'autres évidemment.

Mais il est important de remarquer que les fréquences constatées n'ont pas de valeur en soi. Pour constituer des symptômes interprétables, il est nécessaire de les examiner relativement à d'autres fréquences ; par exemple, celles d'autres événements de même nature dans le même texte, ou bien celles des mêmes événements dans d'autres textes. WEBLEX est conçu pour mettre en œuvre de telles comparaisons. Si l'on se reporte au tableau 3 de

la contribution d'Arnold et Geffroy, on peut évaluer l'importance relative des usages respectifs de « genre humain », « humanité » et « espèce humaine » dans ce corpus : le premier est un peu plus fréquent que le second, lequel est à peu près le double du troisième. De même, on peut lire sur ce tableau que la fréquence 22 de « hommes » dans HESP traduit un suremploi de ce mot dans cet article, tandis que 57 dans ENCY est à interpréter comme un sous-emploi. De tels constats ne vont pas de soi.

Une vision synthétique des caractéristiques lexicales les plus remarquables d'un corpus se constitue ainsi peu à peu. Ces synthèses globales offrent des heuristiques pour interpréter les données. On trouve dans WEBLEX différents outils de classement orientés vers la *spécificité* d'apparition dans une partie plutôt qu'une autre (analyse contrastive, voir tableau 3 dans Arnold et Geffroy), vers la *cooccurrence* d'unités dans les phrases (analyse d'attirances - associations, voir tableau 6 dans Arnold et Geffroy) ou encore vers la *répartition* plus ou moins régulière dans un texte (analyse de progression).

Bien sûr, toutes les interprétations-intuitions suggérées par les constats quantitatifs et l'appréhension globale de tendances au moyen d'heuristiques doivent être affinées et contrôlées au moyen d'outils de lecture et de recherche permettant de res(t)ituer chaque phénomène textuel dans son contexte d'apparition. Ce contrôle correspond à ce que nous appelons l'approche *qualitative* de la méthode. On trouve dans WEBLEX différents outils, fondés sur un langage d'expression algébrique[2], permettant de relocaliser dans leur contexte tous les événements textuels dont la fréquence est apparue comme remarquable et ainsi de les examiner cas par cas. Ce langage permet d'exprimer un événement textuel simple ou complexe (c'est-à-dire exprimé par une séquence lexico-syntaxique) borné par les limites de phrase. L'ensemble des événements recensés est alors lisible sous la forme de concordances KWIC triées, de contextes plus ou moins étendus (voir notamment dans Arnold et Geffroy les tableaux 4, 5 et 7), pouvant aller jusqu'au corpus entier selon l'édition paginée disponible en ligne.

La mise en œuvre de WEBLEX consiste donc en un va-et-vient permanent entre les analyses quantitatives de synthèses et le contrôle local fin qu'offre la lecture assistée. Ce logiciel exploite au maximum l'interface hypertextuelle de l'Internet pour assister l'utilisateur dans ce va-et-vient à l'aide de divers parcours structurés par la donnée et par les actions disponibles.

2. Basé sur le moteur de recherche CQP (pour *Corpus Query Processor*) de l'*Institut für Maschinelle Sprachverarbeitung* de l'Université de Stuttgart (Christ 94).

3. Fiabilité de l'édition électronique

Dans leur contribution Arnold et Geffroy notent quelques défauts qui subsistent dans l'édition électronique de l'*Encyclopédie* : mauvais découpages du texte créant des « articles fantômes », fausses attributions d'auteur..., la liste est certainement incomplète. Mais qu'en est-il de la fidélité typographique du texte ? La question se pose évidemment pour la fiabilité des questionnements documentaires par PHILOLOGIC. Elle se pose avec plus d'acuité encore pour les explorations quantitatives par WEBLEX. C'est ce problème que nous étudions maintenant.

Faute de temps et de moyens, nous avons travaillé sur un tout petit échantillon du texte, 10 pages seulement sur les 16 000 de l'*Encyclopédie* in folio, extrait de deux articles (ÉCONOMIE POLITIQUE et ECLECTISME) du corpus Hugees. Dans ces 10 pages, nous avons relevé systématiquement, par double lecture, les différences entre le fac-similé de l'édition papier de référence et la version texte de l'édition électronique. Pour fournir une présentation claire des résultats, nous avons établi une typologie sommaire des « coquilles ». Trois types ont été distingués :

Les coquilles de graphème, exemples :

p. 11:367, [c->e] (c dans l'édition papier est transcris par e dans l'édition électronique) dans théo[e]ratie, de même [t->r] dans é[r]at.

Les coquilles d'accent, exemples :

p. 11:378, [é->e] dans stérilit[e], p. 11:381, [é-> e] dans sp[e]culations.

Les coquilles de segmentation, exemples :

p. 5:279 [- (tiret de fin de ligne) -> ^ (espace)] dans E^clectisme, de même [-- ->-^-] dans disoit-^-il.

Le tableau 1 présente le nombre de coquilles relevées par type dans chaque page de notre échantillon. La localisation comporte le volume, la page et l'article.

localisation	Nb. coquilles	graphèmes	accents	segmentations
p.11:367 OECO	36	23	05	08
p.11:369 OECO	24	08	07	09
p.11:376 OECO	04	01	01	02
p.11:378 OECO	06	03	01	02
p.11:379 OECO	07	06	01	00
p.11:381 OECO	06	02	01	03
p.11:382 OECO	08	04	00	04
p.5:278 ECLT	20	11	02	07
p.5:279 ECLT	20	05	04	11
p.5:281 ECLT	39	18	02	19
Total	170	81	24	65

Tableau 1 : statistique des coquilles par types

On constate d'emblée que les coquilles, comme on pouvait s'y attendre, sont loin d'être réparties uniformément. On passe de chiffres très faibles, 4 ou 6, à des chiffres relativement importants, 36 ou 39.

Au sein d'une même page, nous avons constaté que les fautes se concentrent souvent au sein d'un même paragraphe. Dans l'échantillon de notre expérience, les passages en latin ou en italique contiennent, en général, beaucoup plus de coquilles de type graphème que le texte courant.

Le tableau 2 permet de voir les coquilles « graphèmes » recensées.

	a	b	c	d	e	f	i	l	m	n	o	r	s	t	u	Ø	ss	rn	oe	tot
a											1				1					2
c					6									2						8
e			1																	1
f												1	7	1						9
h		1														6				7
i												2	1	1						4
l						1	2						1	1						5
n							1								1					2
o	1		1	1																3
p						1														1
q											1									1
r					1		2							4		1				8
s						4	1	2						2						9
t							1					3	2		1					7
u	1									1										2
z																1				1
ï														1						1
ff																	3			3
if															1					1
ni									1											1
nn																		1		1
ae																			4	4
tot	2	1	2	1	7	6	7	2	1	1	2	6	11	12	4	8	3	1	4	81

Tableau 2 : coquilles de graphème

Sur la première ligne du tableau, on peut lire que le graphème *a* a été transcrit une fois en *o* et une fois en *u*, donc qu'il est à l'origine de deux erreurs (colonne tot). Le symbole Ø indique non pas un changement de graphème mais son omission dans l'édition électronique. Par exemple, à la ligne *h*, on lit que ce graphème a été changé une fois en *b*, et par ailleurs, a été omis 6 fois. Les graphèmes absents de la première colonne du tableau (b, d, g, etc.) n'ont donné lieu à aucune coquille dans notre enquête.

On constate, là encore, que les coquilles affectent les lettres de manière assez inégale. On pouvait évidemment s'attendre à ce que les graphèmes *s* et *f* soient les plus affectés à cause de la typographie originale. C'est bien le cas, la dernière colonne du tableau indique que chacun d'eux a subi 9 changements. Mais *c*, *r* et *t* sont également souvent mal transcrits, respectivement 8, 8 et 7 fois. Enfin, *h* est souvent omis, en particulier à l'initiale des mots. Nous reviendrons plus bas sur ce dernier point.

Les tableaux 3 et 4, présentent respectivement le détail des coquilles d'accent et de segmentation.

	e	é	è	ê	a	á	i	Tot
e		1	4					5
é	10			1				11
è	2							2
ê	1		1					2
a						1		1
á					1			1
î							2	2
Tot	13	1	5	1	1	1	2	24

Tableau 3 : coquilles d'accent

	-	^	Ø	Tot
-		19		19
-l		15	6	21
-c		9		9
^	1		8	9
Ø		1		1
a		2		2
i		3		3
'			1	1
Tot	1	49	15	65

Tableau 4 : coquilles de segmentation

Concernant les accents, il y a peu à dire. On voit très bien que la majorité des coquilles touchent les diacritiques du *e*, aigu, grave et circonflexe. Certains *e* ont été fautivement diacrisés[3].

Le tableau 4 contient les coquilles de segmentation, il est plus difficile à lire :

[-] est le trait d'union interne à une graphie,

[-l] est le tiret indiquant une coupure de graphie en fin de ligne,

[-c] est le tiret indiquant une coupure de graphie en fin de première colonne (problème important car l'édition papier, comme on sait, comporte systématiquement deux colonnes),

[^] indique l'espace,

[Ø] indique l'omission comme dans le tableau 2.

Les coquilles de segmentation sont assez nombreuses en regard de celles des graphèmes, 65 par rapport à 81. Elles semblent être plus systématiques que les autres coquilles. Beaucoup des tirets de fin de ligne ont été ôtés dans l'édition électronique, mais ils sont souvent remplacés par des espaces ou fautivement omis (quand un mot est coupé dans l'édition papier sur son trait d'union interne). Les traits d'union internes à des mots marqués par -- sont également très souvent scindés par un espace, enfin les mots longs en fin de colonne sont presque systématiquement coupés en deux (9 fois sur les 10 pages traitées).

Du point de vue du lexique les coquilles de segmentation engendrent beaucoup de graphies fautives. Les deux colonnes [^] et [Ø] du tableau 4 produisent (2 x 49 + 15), soit 113 graphies fantômes.

Quel bilan tirer de cette étude ? D'abord qu'il faudrait l'étendre pour avoir des conclusions plus solides, l'échantillon retenu est trop petit pour être représentatif d'un texte d'une telle ampleur. Mais 100 à 200 pages prises au hasard pourraient suffire, nous semble-t-il, pour avoir des conclusions sérieuses sur la fiabilité typographique du texte électronique de l'*Encyclopédie*. Ce n'est pas un très gros travail, il est même négligeable en regard de celui qui a déjà été accompli pour numériser le texte. Une telle étude permettrait d'évaluer le volume des corrections à entreprendre et d'orienter celles-ci en vue d'une meilleure efficacité. En attendant donc une étude plus complète qui reste à faire, notre petit travail conduit à estimer provisoirement le taux d'erreur dans le texte électronique à 170/10 (voir tableau 2), soit 17 coquilles en moyenne par page.

Il nous semble que pour les explorations textuelles documentaires de PHILOLOGIC le taux d'erreur constaté est tout à fait acceptable en général. Les mesures lexicométriques, en revanche, faites sur ces textes pâtissent sans doute de la qualité de l'enregistrement.

3. La variabilité d'usage des signes diacritiques d'une même graphie est connue au XVIII^e^ siècle. Ce n'est pas ce qui nous intéresse ici. Seule compte une infidélité à l'édition de référence.

Dans notre enquête sur les coquilles de type graphème, un cas cependant a attiré notre attention, celui de la disparition du *h* (voir tableau 2, ligne h). Nous avions constaté, dans l'échantillon choisi, que *h* était omis 6 fois, chaque fois à l'initiale du mot « humaine ».

Nous avons donc fait des interrogations à l'aide de PHILOLOGIC sur plusieurs mots, notamment sur des noms propres commençant par *H* : Homère, Hésiode, Horace, Hippocrate.

Un extrait du résultat de cette interrogation constitue le tableau 5 ci-après.

ENCYCLOPÉDIE OU *DICTIONNAIRE RAISONNÉ* DES SCIENCES, DES ARTS ET DES MÉTIERS

Critères bibliographiques: **none** (**Toutes** documents) [*sic*]
Critères de recherche: **EsiodeloracelomErelippocrate**
Votre recherche a trouvé **308** occurrences

Occurrences 1-308:

1. ACCIDENT (1:72)on début. C'est ainsi qu'**orace** a dit au commencement d'une Ode:
2. ACTE (1:116)le soit en cinq actes, & qu'**orace** ait eu raison d'en faire un préce
3. Acteur (1:117)teurs à la fois: regle qu'**orace** a exprimée dans ce vers, Nec quar
4. AMATHRE (1:317)MATHRE * AMATHRE, nom qu'**omere** a donné à une des cinquante Néréi
5. AMPHINOME (1:376)ME * AMPHINOME, nom qu'**omere** donne à une des cinquante Néréide
6. ANATOMIE (1:411)ve l'Anatomie; & lorsqu'**ippocrate** fut appellé par les Abderitai
7. ANATOMIE (1:411)etote confond, ainsi qu'**ippocrate**, les nerfs, les ligamens & le
8. ANATOMIE (1:412)it mieux ces parties qu'**ippocrate** ne les avoit connues; & que l
9. ANNEAU (1:479) qu'il en donne, c'est qu'**omere** n'en fait point mention: mais que
(.....)
100. GOUTTE (7:773) rien de limite, quoiqu'**ippocrate**, aphorisme xlix. sect. 6. la
101. Grece (7:806)e: effets merveilleux qu'**orace** a peints avec force, & Ovide avec
102. Grece (7:808)e Crete, &c. Voilà ce qu'**ésiode** nous a transmis en très--beaux v
103. Grece (7:808)lontiere. L'un disoit qu'**omere** n'avoit pas vingt ans à être lu:
104. Grotte du Chien (7:968)si ancienne qu'**omere**: car le mont Arima où il place ce
105. GYMNIQUES (7:1019)marque Eustathe, qu'**omere**, grand observateur des bienséan c
106. HAMADRIADE (8:33)bres. Je sai bien qu'**ésiode** donne à leur vie une durée prodi
107. HECAERGUE, o... (8:93)1.) épithete qu'**omere** donne souvent à Apollon, à Diane,
108. HELLOPES (8:107) le fertile canton qu'**ésiode** nomme Hellopie, n'étoit autre ch
109. HÉMISTICHE (8:124)-près semblable, qu'**orace** les imita quelquefois lorsque le
110. HÉMORRHAGIE (8:118)ntité, c'est ce qu'**ippocrate** appelle EN)R(USIN, ou S2ALAGM
(.....)
300. Visage (17:337)ippocratique, parce qu'**ippocrate** en a fait la peinture suivant
301. UNGUENTARIUS (17:380)est pour cela qu'**orace** appelle les parfumeurs, tusci tur
302. Vomissement ... (17:467)isons, quoiqu'**ippocrate** ait excepté l'hiver, quoiqu'i
303. Vomissement ... (17:467)hiver, quoiqu'**ippocrate** ait exclus cette saisons; & da
304. Voyageur (17:478) voyage. C'est ce qu'**orace** & Virgile appellent vota vestes.
305. Urine (17:500)rvation: c'est ainsi qu'**ippocrate** l'a traitée, & qu'il convient
306. Urine (17:512)uasi dans ce viscere qu'**ippocrate** en marque l'origine. Lorsqu'i
307. Zéphire (17:706)étoit un des vents qu'**ésiode** dit etre enfans des dieux. Anchi
308. CONTRAT (17:766)e nos contrats, ce qu'**orace** disoit de ceux de son tems. Adde

Tableau 5 : extrait d'une interrogation sur quatre noms propres avec *H* à l'initiale

On trouve 308 occurrences erronées dans lesquelles le *H* initial est omis dans le texte. Il est peu vraisemblable que ces omissions proviennent d'erreurs aléatoires. La même interrogation avec le *H* à l'initiale renvoie 2 432 occurrences. Une telle interrogation documentaire génèrerait donc plus de 10 % de silence. Il n'est pas impossible que d'autres coquilles systématiques se soient glissées dans l'édition électronique. Seule une enquête plus approfondie supprimerait les doutes qui peuvent subsister[4].

La numérisation des grandes œuvres du passé progresse très rapidement[5]. D'ores et déjà, sur le seul site de l'ATILF à Nancy (http://www.inalf.fr/atilf/) sont consultables, outre l'*Encyclopédie*, plusieurs grands dictionnaires, le *Dictionarium latinogallicum* de Robert Estienne, le *Dictionnaire historique et critique* de Pierre Bayle, et plusieurs éditions du *Dictionnaire de l'Académie française*. On peut ne voir dans ce mouvement qu'un nouveau mode de conservation et de transmission de ces textes et un moyen de les rendre disponibles à un public beaucoup plus large. Mais il ne s'agit pas seulement de cela.

Au-delà d'un changement de support, l'informatique assiste l'utilisateur et lui donne des moyens d'investigation, d'extraction et d'analyse, qui apparaissent tout à la fois comme très rudimentaires et très considérables. Il est clair que la familiarité avec ces nouveaux instruments permet de poser aux œuvres des questions inédites, suggère et rend possible des analyses auxquelles personne n'avait jamais songé auparavant. C'est à travers ces pratiques que s'élaborent peu à peu de nouveaux modes de lecture, de nouvelles compétences et de nouvelles perceptions. Nous avons tenté de montrer qu'il est nécessaire de distinguer deux niveaux dans ces pratiques, celui de l'enregistrement ou encodage de l'objet textuel et celui des commandes d'actions possibles sur cet objet. Ces deux niveaux se confondent trop souvent dans une seule boite noire impénétrable.

D'un côté, il est important de veiller à ce que les objets textuels soient fidèlement enregistrés. Cela va, bien sûr, très au delà de la correction orthographique, et suppose un encodage qui respecte les diverses entités sémantiquement différenciées dans le texte et permette leur reconnaissance automatique pour être mises en œuvre dans les requêtes. Les limites de ces entités ne sont malheureusement pas toujours marquées par des distinctions typographiques dans l'édition papier.

D'un autre côté, même si c'est techniquement difficile à réaliser, il faudrait donner à l'utilisateur le libre choix entre plusieurs logiciels

4. Voir sur ce point la communication de Robert Morrissey, p. 283, dans le même numéro.

5. Nous renvoyons sur ce point au livre de Jean Pruvost (2000) très bien documenté.

(PHILOLOGIC, WEBLEX et beaucoup d'autres qui existent) pour formuler ses requêtes. L'utilisateur pourrait ainsi choisir le plus adapté à ses travaux, et acquérir une distance critique et des exigences dans son dialogue avec la machine.

Serge HEIDEN
Pierre LAFON
UMR 8503, CNRS et ENS-LSH
« Analyses de corpus linguistiques, usages, traitements »

Références

Christ, Oliver (1994), *A modular and flexible architecture for an integrated corpus query system* in Proceedings of COMPLEX'94: 3rd Conference on Computational Lexicography and Text Research, pp. 23-32, Budapest, Hungary cmp-lg: 9408005.

Lafon P. (1984), *Dépouillements et statistiques en lexicométrie*, Slatkine-Champion, Genève-Paris.

Lebart L., Salem A. (1994), *Statistique Textuelle*, Dunod, Paris.

Pruvost J. (2000), *Dictionnaires et nouvelles technologies*, Écritures électroniques, PUF, Paris.